On the conditions used to prove oracle results for the Lasso

Sara van de Geer & Peter Bühlmann
ETH Zürich

September, 2009

Abstract

Oracle inequalities and variable selection properties for the Lasso in linear models have been established under a variety of different assumptions on the design matrix. We show in this paper how the different conditions and concepts relate to each other. The restricted eigenvalue condition (Bickel et al., 2009) or the slightly weaker compatibility condition (van de Geer, 2007) are sufficient for oracle results. We argue that both these conditions allow for a fairly general class of design matrices. Hence, optimality of the Lasso for prediction and estimation holds for more general situations than what it appears from coherence (Bunea et al., 2007b,c) or restricted isometry (Candès and Tao, 2005) assumptions.

Keywords and phrases: Coherence, compatibility, irrepresentable condition, Lasso, restricted eigenvalue, restricted isometry, sparsity.

1 Introduction

In this paper we revisit some sufficient conditions for oracle inequalities for the Lasso in regression and examine their relations. Such oracle results have been derived, among others, by Bunea et al. (2007c), van de Geer (2008), Zhang and Huang (2008), Meinshausen and Yu (2009), Bickel et al. (2009), and for the related Dantzig selector by Candès and Tao (2007) and Koltchinskii (2009b). Furthermore, variable selection properties of the Lasso have been studied by Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Lounici (2008), Zhang (2009) and Wainwright (2009). Our main aim is to present an overview of the relations (of which some are known and some are new), and to emphasize that that sufficient conditions for oracle inequalities hold in fairly general situations.

The Lasso, which we at first only study in a noiseless situation, is defined as follows. Let $\mathcal{X}$ be some measurable space, Q be a probability measure on $\mathcal{X}$, and $\|\cdot\|$ be the $L_2(Q)$ norm. Consider a fixed dictionary of functions $\{\psi_j\}_{j=1}^p \subset L_2(Q)$, and linear functions

$$f_\beta(\cdot) := \sum_{j=1}^p \beta_j \psi_j(\cdot) : \ \beta \in \mathbb{R}^p.$$

Consider moreover a fixed target

$$f^0(\cdot) := \sum_{j=1}^p \beta_j^0 \psi_j(\cdot).$$

We let $S := \{j : \ \beta_j^0 \neq 0\}$ be its active set, and $s := |S|$ be the *sparsity index* of f^0.

For some fixed $\lambda > 0$, the Lasso for the noiseless problem is

$$\beta^* := \arg\min_\beta \left\{ \|f_\beta - f^0\|^2 + \lambda\|\beta\|_1 \right\}, \tag{1}$$

where $\|\cdot\|_1$ is the ℓ_1-norm. We write $f^* := f_{\beta^*}$ and let S_* be the active set of the Lasso.

Let us precise what we mean by an oracle inequality. With β being a vector in $\mathbb{R}^p$, and $\mathcal{N} \subset \{1, \ldots, p\}$ an index set, we denote by

$$\beta_{j,\mathcal{N}} := \beta_j \mathrm{l}\{j \in \mathcal{N}\},\ j = 1, \ldots, p,$$

the vector with non-zero entries in the set $\mathcal{N}$ (hence, for example $\beta^0_S = \beta^0$).

Definition: Sparsity constant and sparsity oracle inequality. *The* sparsity constant ϕ_0 *is the largest value* $\phi_0 > 0$ *such that Lasso with* β^* *and* f^* *satisfies the* ϕ_0-sparsity oracle inequality

$$\|f^* - f^0\|^2 + \lambda\|\beta^*_{S^c}\|_1 \leq \frac{\lambda^2 s}{\phi_0^2}.$$

Restricted eigenvalue conditions (see Koltchinskii (2009a,b) and Bickel et al. (2009)) have been developed to derive lower bounds for the sparsity constant. We will present these conditions in the next section. Irrepresentable conditions (see Zhao and Yu (2006)) are tailored for proving variable selection, i.e., showing that $S_* = S$, or, more more modestly, that the symmetric difference $S_* \triangle S$ is small.

1.1 Organization of the paper

We start out with, in Section 2, an overview of the conditions we will compare, and some pointers to the literature. Once the conditions are made explicit, we give in Subsection 2.2 a summary of the various relations. Figure 1 displayed there enables to see these at a single glance. We give a proof of each of the indicated (numbered) implications. Sections 3 - 9 rigorously deal with all the different cases. The weakest condition is a compatibility condition. Stronger conditions can rule out many interesting cases. We illustrate in Section 10 that one may check compatibility using approximations. We give several examples, where the compatibility condition holds. We also give an example where the compatibility condition yields a major improvement to the oracle result, as compared to the restricted eigenvalue condition. The noisy case, studied briefly in Section 11, poses no additional theoretical difficulties. A lower bound on the regularization parameter λ is required, and implications become somewhat more technical because all further results depend on this lower bound. Section 12 discusses the results.

1.2 Some notation

For a vector v, we invoke the usual notation

$$\|v\|_q = \begin{cases} (\sum_j |v_j|^q)^{1/q}, & 1 \leq q < \infty \\ \max_j |v_j|, & q = \infty \end{cases} .$$

The Gram matrix is

$$\Sigma := \int \psi^T \psi dQ,$$

so that

$$\|f_\beta\|^2 = \beta^T \Sigma \beta.$$

The entries of Σ are denoted by $\sigma_{j,k} := (\psi_j, \psi_k)$, with $(\cdot, \cdot)$ being the inner product in $L_2(Q)$.

To clarify the notions we shall use, consider for a moment a partition of the form

$$\Sigma := \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix},$$

where $\Sigma_{1,1}$ is an $N \times N$ matrix, $\Sigma_{2,1}$ is a $(p-N) \times N$ matrix and $\Sigma_{1,2} := \Sigma_{2,1}^T$ is its transpose, and where $\Sigma_{2,2}$ is a $(p-N) \times (p-N)$ matrix. Such partitions will be play an important role in the sections to come.

More generally, for a set $\mathcal{N} \subset \{1, \ldots, p\}$ with size N, we introduce the $N \times N$ matrix

$$\Sigma_{1,1}(\mathcal{N}) := (\sigma_{j,k})_{j,k \in \mathcal{N}},$$

the $(p-N) \times N$ matrix

$$\Sigma_{2,1}(\mathcal{N}) = (\sigma_{j,k})_{j \notin \mathcal{N}, k \in \mathcal{N}},$$

and the $(p-N) \times (p-N)$ matrix

$$\Sigma_{2,2}(\mathcal{N}) := (\sigma_{j,k})_{j,k \notin \mathcal{N}}.$$

We let $\Lambda^2_{\min}(\Sigma_{1,1}(\mathcal{N}))$ be the smallest eigenvalue of $\Sigma_{1,1}(\mathcal{N})$. Throughout, we assume that, for the fixed active set S, the smallest eigenvalue $\Lambda^2_{\min}(\Sigma_{1,1}(S))$ is strictly positive, i.e., that $\Sigma_{1,1}(S)$ is non-singular.

We sometimes identify $\beta_{\mathcal{N}}$ with the vector $|\mathcal{N}|$-dimensional vector $\{\beta_j\}_{j \in \mathcal{N}}$, and write e.g.,

$$\beta_{\mathcal{N}}^T \Sigma \beta_{\mathcal{N}} = \beta_{\mathcal{N}}^T \Sigma_{1,1}(\mathcal{N}) \beta_{\mathcal{N}}.$$

2 An overview of definitions

The definitions we will present are conditions on the Gram matrix Σ, namely conditions on quadratic forms $\beta^T \Sigma \beta$, where β is restricted to lie in some subset of $\mathbb{R}^p$. We first take the set of restrictions

$$\mathcal{R}(L, S) := \{\beta : \|\beta_{S^c}\|_1 \le L \|\beta_S\|_1 \ne 0\}.$$

The compatibility condition we discuss here is from van de Geer (2007). Its name is based on the idea that we require the ℓ_1-norm and the $L_2(Q)$-norm to be somehow compatible.

Definition: Compatibility condition. *We call*

$$\phi^2_{\text{compatible}}(L, S) := \min\left\{ \frac{s\|f_\beta\|^2}{\|\beta_S\|_1^2} : \ \beta \in \mathcal{R}(L, S) \right\}$$

the (L,S)-restricted ℓ_1-eigenvalue.
The (L,S)-compatibility condition *is satisfied if* $\phi_{\rm compatible}(L,S)>0$.

The bound $\|\beta_S\|_1 \le \sqrt{s}\|\beta_S\|_2$ (which holds for any β) leads to two successively stronger versions of restricted eigenvalues. We moreover consider supsets $\mathcal{N}$ of S with size at most N. Throughout in our definitions, $N \ge s$. We will only invoke $N=s$ and $N=2s$ (for simplicity).

Define the sets of restrictions

$$\mathcal{R}_{\rm adaptive}(L,S) := \{\beta :\ \|\beta_{S^c}\|_1 \le \sqrt{s}L\|\beta_S\|_2\},$$

and for $\mathcal{N}\supset S$,

$$\mathcal{R}(L,S,\mathcal{N}) := \{\beta \in \mathcal{R}(L,S) :\ \|\beta_{\mathcal{N}^c}\|_\infty \le \min_{j\in \mathcal{N}\backslash S} |\beta_j|\},$$

and

$$\mathcal{R}_{\rm adaptive}(L,S,\mathcal{N}) := \{\beta \in \mathcal{R}_{\rm adaptive}(L,S) :\ \|\beta_{\mathcal{N}^c}\|_\infty \le \min_{j\in \mathcal{N}\backslash S} |\beta_j|\}.$$

If $N=s$, we necessarily have $\mathcal{N}\backslash S=\emptyset$. In that case, we let $\min_{j\in\mathcal{N}\backslash S}|\beta_j|=0$, i.e., $\mathcal{R}(L,S,S)=\mathcal{R}(L,S)$ ($\mathcal{R}_{\rm adaptive}(L,S,S)=\mathcal{R}_{\rm adaptive}(L,S)$).

The restricted eigenvalue condition is from Bickel et al. (2009) and Koltchinskii (2009b). We complement it with the *adaptive* restricted eigenvalue condition. The name of the latter is inspired by the fact that this strengthened version is useful for the development of theory for the *adaptive* Lasso (Zou, 2006) which we do not show in this paper.

Definition: (Adaptive) restricted eigenvalue. *We call*

$$\phi^2(L,S,N) := \min\left\{ \frac{\|f_\beta\|^2}{\|\beta_{\mathcal{N}}\|_2^2} :\ \mathcal{N}\supset S,\ |\mathcal{N}|\le N,\ \beta\in\mathcal{R}(L,S,\mathcal{N})\right\}$$

the (L,S,N)-restricted eigenvalue*, and, similarly,*

$$\phi^2_{\rm adaptive}(L,S,N) := \min\left\{ \frac{\|f_\beta\|^2}{\|\beta_{\mathcal{N}}\|_2^2} :\ \mathcal{N}\supset S,\ |\mathcal{N}|\le N,\ \beta\in\mathcal{R}_{\rm adaptive}(L,S,\mathcal{N})\right\}$$

the adaptive (L,S,N)-restricted eigenvalue. *The* (adaptive) (L,S,N)-restricted eigenvalue condition *holds if* $\phi(L,S,N)>0$ *(*$\phi_{\rm adaptive}(L,S,N)>0$*)* .

We introduce the (adaptive) restricted regression condition to clarify various connections between different assumptions.

Definition: (Adaptive) restricted regression. *The* (L,S,N)-restricted regression *is*

$$\vartheta(L,S,N) := \max\left\{ \frac{|(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{N}^c}})|}{\|f_{\beta_{\mathcal{N}}}\|^2} :\ \ \mathcal{N}\supset S,\ |\mathcal{N}|\le N,\ \beta\in\mathcal{R}(L,S,\mathcal{N})\right\}.$$

The adaptive (L,S,N)-restricted regression *is*

$$\vartheta_{\text{adaptive}}(L,S,N) :=$$
$$\max\left\{ \frac{|(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{N}^c}})|}{\|f_{\beta_{\mathcal{N}}}\|^2} : \ \mathcal{N} \supset S,\ |\mathcal{N}| \le N,\ \beta \in \mathcal{R}_{\text{adaptive}}(L,S,\mathcal{N}) \right\}.$$

The (adaptive) (L,S,N)-restricted regression condition *holds if* $\vartheta(L,S,N) < 1$ *(*$\vartheta_{\text{adaptive}}(L,S,N) < 1$*).*

Note that $(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{N}^c}})/\|f_{\beta_{\mathcal{N}}}\|^2$ equals the coefficient when regressing $f_{\beta_{\mathcal{N}^c}}$ onto $f_{\beta_{\mathcal{N}}}$.

Of course all these definitions depend on the Gram matrix Σ. In Sections 10 and 11, we make this dependence explicit by adding the argument Σ, e.g. the (Σ, L, S)-compatibility condition, etc.

When $L = 1$, the argument L is omitted, e.g. $\phi_{\text{compatible}}(S) := \phi_{\text{compatible}}(1, S)$, and e.g., the S-compatibility condition is then the condition $\phi_{\text{compatible}}(S) > 0$. The case $L > 1$ is mainly needed to handle the situation with noise, and $L < 1$ is of interest when studying the *adaptive* Lasso (but we do not develop its theory in this paper).

We now present some definitions from Candès and Tao (2005).

Definition: Restricted orthogonality constant. *The quantity*

$$\theta(S,N) := \sup_{\mathcal{N} \supset S:\ |\mathcal{N}| \le N} \ \sup_{\mathcal{M} \subset \mathcal{N}^c,\ |\mathcal{M}| \le s} \ \sup_{\beta} \left| \frac{(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{M}}})}{\|\beta_{\mathcal{N}}\|_2 \|\beta_{\mathcal{M}}\|_2} \right|,$$

is called the (S,N)-restricted orthogonality constant. *We moreover define*

$$\theta_{s,N} := \max\{\theta(S,N) :\ |S| = s\}.$$

Definition: Restricted isometry constant. *The* N-restricted isometry constant *is the smallest value of* δ_N *such that for all* $\mathcal{N}$ *with* $|\mathcal{N}| \le N$*,*

$$(1 - \delta_N)\|\beta_{\mathcal{N}}\|_2^2 \le \|f_{\beta_{\mathcal{N}}}\|^2 \le (1 + \delta_N)\|\beta_{\mathcal{N}}\|_2^2.$$

Definition: Uniform eigenvalue. *The* (S,N)*-uniform eigenvalue is*

$$\Lambda^2(S,N) := \inf_{\mathcal{N} \supset S,\ |\mathcal{N}| \le N} \Lambda^2_{\min}(\Sigma_{1,1}(\mathcal{N})).$$

As mentioned before, we always assume that $\Lambda(S,s) > 0$.

Definition: Weak restricted isometry. *The* weak (S,N)-restricted isometry constant *is*

$$\vartheta_{\text{weak-RIP}}(S,N) := \frac{\theta(S,N)}{\Lambda^2(S,N)}.$$

The weak (L,S,N)-restricted isometry property *holds if* $\vartheta_{\text{weak-RIP}}(S,N) < 1/L$*.*

Definition: Restricted isometry property. *The* RIP constant *is*

$$\vartheta_{\text{RIP}} := \frac{\theta_{s,2s}}{1 - \delta_s - \theta_{s,s}}.$$

The restricted isometry property*, shortly* RIP*, holds if* $\vartheta_{\text{RIP}} < 1$.

An irrepresentable condition can be found in Zhao and Yu (2006). We use a modified version which involves only the design but not the true coefficient vector β^0 (whereas its sign vector appears in Zhao and Yu (2006)). The reason is that most other conditions considered in this paper do not depend on β^0 as well. Our (L, S, N)-irrepresentable condition with $L = 1$ and $N = s$ is only slightly stronger than the condition in Zhao and Yu (2006).

Definition: Irrepresentable condition.
Part 1. *We call*

$$\vartheta_{\text{irrepresentable}}(S, N) := \min_{\mathcal{N} \supset S:\ |\mathcal{N}| \le N} \max_{\|\tau_{\mathcal{N}}\|_\infty \le 1} \|\Sigma_{2,1}(\mathcal{N})\Sigma_{1,1}^{-1}(\mathcal{N})\tau_{\mathcal{N}}\|_\infty$$

the (S, N)-uniform irrepresentable constant*. The* (L, S, N)-uniform irrepresentable condition *is met, if* $\vartheta_{\text{irrepresentable}}(S, N) < 1/L$.
Part 2. *We say that the* (L, S, N)-irrepresentable condition *is met, if for some* $\mathcal{N} \supset S$ *with* $|\mathcal{N}| \le N$*, and all vectors* $\tau_{\mathcal{N}}$ *satisfying* $\tau_{\mathcal{N}} \in \{-1, 1\}^{|\mathcal{N}|}$*, we have*

$$\|\Sigma_{2,1}(\mathcal{N})\Sigma_{1,1}^{-1}(\mathcal{N})\tau_{\mathcal{N}}\|_\infty < 1/L.$$

Part 3. *We say that the* weak (S, N)-irrepresentable condition *is met, if for all* $\tau_S \in \{-1, 1\}^s$*, and for some* $\mathcal{N} \supset S$ *with* $|\mathcal{N}| \le N$*, and for some* $\tau_{\mathcal{N} \setminus S} \in \{-1, 1\}^{|\mathcal{N} \setminus S|}$*, we have*

$$\|\Sigma_{2,1}(\mathcal{N})\Sigma_{1,1}^{-1}(\mathcal{N})\tau_{\mathcal{N}}\|_\infty \le 1.$$

Finally, we present coherence conditions, which are in the spirit of Bunea et al. (2007b,c). Cai et al. (2009b) derive an oracle result under a tight coherence condition.

Definition: Coherence. *The* (L, S)-mutual coherence condition *holds if*

$$\vartheta_{\text{mutual}}(S) := \frac{s \max_{j \notin S} \max_{k \in S} |\sigma_{j,k}|}{\Lambda^2(S, s)} < 1/L.$$

The (L, S)-cumulative coherence condition *holds if*

$$\vartheta_{\text{cumulative}}(S) := \frac{\sqrt{s}\sqrt{\sum_{k \in S}\left(\sum_{j \notin S} |\sigma_{j,k}|\right)^2}}{\Lambda^2(S, s)} < 1/L.$$

2.1 Implications for the Lasso and some first relations

It is shown in van de Geer (2007) that the compatibility condition implies oracle inequalities for the Lasso. We re-derive the result for later reference and also for illustrating that the compatibility condition is just a condition to make the proof go through. We also show (again for later reference) the additional ℓ_2-result if one uses the (S, N)-restricted eigenvalue condition.

Lemma 2.1 *(Oracle inequality) We have for the Lasso in (1),*

$$\|f^* - f^0\|^2 + \lambda \|\beta^*_{S^c}\|_1 \le \lambda^2 s / \phi^2_{\text{compatible}}(S).$$

Moreover, letting $\mathcal{N}_* \backslash S$ *being the set of the* $N - s$ *largest coefficients* $|\beta^*_j|$, $j \in S^c$,

$$\|\beta^*_{\mathcal{N}_*} - \beta^0_{\mathcal{N}_*}\|_2^2 \le \lambda^2 s / \phi^4(S, N).$$

Proof of Lemma 2.1. The first assertion follows from the Basic Inequality

$$\|f^* - f^0\|^2 + \lambda \|\beta^*\|_1 \le \lambda \|\beta^0\|_1,$$

using the definition of the Lasso in (1), which implies

$$\|f^* - f^0\|^2 + \lambda \|\beta^*_{S^c}\|_1 \le \lambda \Big(\|\beta^0\|_1 - \|\beta^*_S\|_1 \Big)$$

$$\le \lambda \|\beta^*_S - \beta^0_S\|_1 \le \lambda \sqrt{s} \|f^* - f^0\| / \phi_{\text{compatible}}(S).$$

Note that the last inequality holds because $\beta^* - \beta^0 \in \mathcal{R}(S)$ which follows by its preceding inequality:

$$\|\beta^*_{S^c}\|_1 = \|\beta^*_{S^c} - \beta^0_{S^c}\|_1 \le \|\beta^*_S - \beta^0_S\|_1.$$

The second result follows from

$$\|\beta^*_{\mathcal{N}_*} - \beta^0_{\mathcal{N}_*}\|_2^2 \le \|f^* - f_0\|^2 / \phi^2(S, N),$$

and using $\phi_{\text{compatible}}(S) \ge \phi(S, N)$. $\square$

An implication of Lemma 2.1 is an ℓ_1-norm result:

$$\begin{aligned} \|\beta^* - \beta^0\|_1 &= \|\beta^*_{S^c}\|_1 + \|\beta^*_S - \beta^0_S\|_1 \\ &\le \lambda s / \phi^2_{\text{compatible}}(S) + \lambda \sqrt{s} \|f^* - f^0\| / \phi_{\text{compatible}}(S) \\ &\le 2 \lambda s / \phi^2_{\text{compatible}}(S), \end{aligned}$$

where the last inequality is using the first assertion in Lemma 2.1. We also note that the second assertion in Lemma 2.1 has most statistical importance for the case with $N = s$. We will need the case $N = 2s$ later in our proofs.

Meinshausen and Bühlmann (2006) and Zhao and Yu (2006) prove that the irrepresentable condition is sufficient and essentially necessary for variable selection, i.e., for achieving

$S_* = S$. We will also present a self-contained proof in Section 6 where we will show that the (S,s)-irrepresentable condition is sufficient and the weak (S,s)-irrepresentable condition is essentially necessary for variable selection.

Bickel et al. (2009) prove oracle inequalities under the restricted eigenvalue condition. They assume

$$\min\{\phi(L,S,s) : \ |S| = s\} > 0$$

(where L can be taken equal to one in the noiseless case).

The restricted isometry property from Candès and Tao (2005), abbreviated to RIP, also requires uniformity in S. They assume the RIP

$$\vartheta_{\rm RIP} < 1.$$

They show that the RIP implies exact reconstruction of β^0 from f^0 by linear programming (that is, by minimizing $\|\beta\|_1$ subject to $\|f_\beta - f^0\| = 0$). Cai et al. (2009a) prove this result assuming $\delta_N + \theta_{s,N} < 1$ for $N = 1.25s$ only; see also Cai et al. (2009) for an earlier result. It is clear that $1 - \delta_N \le \Lambda^2(S,N)$, i.e., the restricted isometry constants are more demanding than uniform eigenvalues. Candès and Tao (2005) furthermore show that

$$\vartheta_{\rm weak-RIP}(S,N) \le \vartheta_{\rm RIP}.$$

See also Figure 1. They prove that the RIP is sufficient for establishing oracle inequalities for the Dantzig selector. Koltchinskii (2009a) and Bickel et al. (2009) show that

$$\phi(L,S,2s) \ge (1 - L\vartheta_{\rm weak-RIP}(S,2s))\Lambda(S,2s).$$

Thus, the weak $(S,2s)$-restricted isometry property implies the $(S,2s)$-restricted eigenvalue condition. See also Figure 1.

Bunea et al. (2007a,b,c) show that their coherence conditions imply oracle results and refinements (see also Section 4 for their condition on the diagonal of Σ). Candès and Plan (2009) weaken the coherence conditions by restricting the parameter space for the regression coefficient β.

Finally, it is clear that $\phi_{\rm adaptive}(L,S,N) \le \phi(L,S,N) \le \phi_{\rm compatible}(L,S)$, i.e.,

adaptive restricted eigenvalue condition $\Rightarrow$

restricted eigenvalue condition $\Rightarrow$

compatibility condition.

See also Figure 1.

It is easy to see that $\vartheta(L,S,N)$ and $\vartheta_{\rm adaptive}(L,S,N)$ scale with L, i.e., we have

$$\vartheta(L,S,N) = L\vartheta(S,N), \ \vartheta_{\rm adaptive}(L,S,N) = L\vartheta_{\rm adaptive}(S,N).$$

This is not true for the (adaptive) restricted (ℓ_1-)eigenvalues. It indicates that the (adaptive) restricted regression is not well-calibrated for proving compatibility or restricted

eigenvalue conditions, i.e, one might pay a large price for taking the route to oracle results via restricted regression conditions.

We end this subsection with the following lemma, which is based on ideas in Candès and Tao (2007). A corollary is the ℓ_2-bound given in (2), which thus illustrates that considering supsets $\mathcal{N}$ of S can be useful. However, we use the lemma for other purposes as well.

We let for any β, $r_j(\beta) := \text{rank}(|\beta_j|)$, $j \in S^c$, if we put the coefficients in decreasing order. Let $\mathcal{N}_0(\beta)$ be the set of the s largest coefficients in S^c:

$$\mathcal{N}_0(\beta) := \{j : \ r_j(\beta) \in \{1, \ldots, s\}\}.$$

Put $\mathcal{N}(\beta) := \mathcal{N}_0(\beta) \cup S$. Further, assuming without loss of generality that $p = (K+2)s$ for some integer $K \geq 0$, we let for $k = 1, \ldots, K$,

$$\mathcal{N}_k(\beta) := \biggl\{ j : \ r_j(\beta) \in \{ks + 1, \ldots, (k+1)s\} \biggr\}.$$

We further define

$$\mathcal{N}_* := \mathcal{N}(\beta^*), \ \mathcal{N}_k^* := \mathcal{N}_k(\beta^*), \ k = 0, 1, \ldots, K.$$

Lemma 2.2 *We have for any any $r \geq 1$, and $1/r + 1/q = 1$, and any β, and for $\mathcal{N} := \mathcal{N}(\beta)$, and $\mathcal{N}_k := \mathcal{N}_k(\beta)$, $k = 0, 1, \ldots, K$, the bound*

$$\|\beta_{\mathcal{N}^c}\|_r \leq \sum_{k=1}^{K} \|\beta_{\mathcal{N}_k}\|_r \leq \|\beta_{S^c}\|_1 / s^{1/q}.$$

Corollary 2.1 *Combining Lemma 2.1 with Lemma 2.2 gives*

$$\|\beta^* - \beta^0\|_2^2 \leq 2\lambda^2 s / \phi^4(S, 2s). \tag{2}$$

This result is from Bickel et al. (2009). The proof we give is essentially the same as theirs.

Proof of Lemma 2.2. Clearly,

$$\|\beta_{\mathcal{N}^c}\|_r = \|\sum_{k=1}^{K} \beta_{\mathcal{N}_k}\|_r \leq \sum_{k=1}^{K} \|\beta_{\mathcal{N}_k}\|_r.$$

We know that for $k = 1, \ldots, K$,

$$|\beta_j| \leq \|\beta_{\mathcal{N}_{k-1}}\|_1 / s, \ j \in \mathcal{N}_k,$$

and hence,

$$\|\beta_{\mathcal{N}_k}\|_r^r \leq s^{-(r-1)} \|\beta_{\mathcal{N}_{k-1}}\|_1^r.$$

It follows that

$$\sum_{k=1}^{K} \|\beta_{\mathcal{N}_k}\|_r \leq \sum_{k=1}^{K} \|\beta_{\mathcal{N}_{k-1}}\|_1 s^{-(r-1)/r} = \|\beta_{S^c}\|_1 / s^{1/q}.$$

□

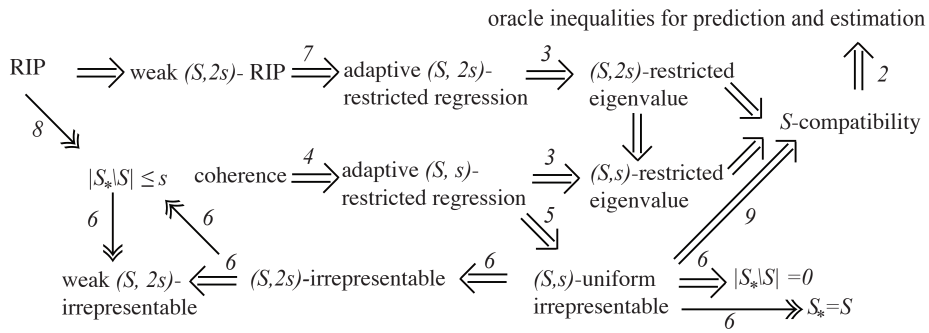

Figure 1: A double arrow ($\Rightarrow$) indicates a straight implication, whereas the more fancy arrowheads mean that the relation is under side-conditions. The numbers indicate the section where the result is (re)proved.

2.2 Summary of the results

The following figure summarizes the results.

Our conclusion is that (perhaps not surprising) the compatibility condition is the least restrictive, and that many sufficient conditions for compatibility may be somewhat too harsh (see also our discussion in Section 12).

3 The restricted regression condition implies the restricted eigenvalue condition

We start out with an elementary lemma.

Lemma 3.1 *Let f_1 and f_2 by two functions in $L_2(P)$. Suppose for some* $0 < \vartheta < 1$.

$$-(f_1, f_2) \leq \vartheta \|f_1\|^2.$$

Then

$$(1 - \vartheta)\|f_1\| \leq \|f_1 + f_2\|.$$

Proof. Write the projection of f_2 on f_1 as

$$f_{2,1}^{\mathrm{P}} := (f_2, f_1)/\|f_1\|^2 f_1.$$

Similarly, let

$$f = (f_1 + f_2)_1^{\mathrm{P}} := (f, f_1)/\|f_1\|^2 f_1$$

be the projection of $f_1 + f_2$ on f_1. Then

$$(f_1 + f_2)_1^{\mathrm{P}} = f_1 + f_{2,1}^{\mathrm{P}} = \left(1 + (f_2, f_1)/\|f_1\|^2\right) f_1,$$

so that

$$\|(f_1+f_2)_1^{\rm P}\| = \left|1+(f_2,f_1)/\|f_1\|^2\right| \|f_1\|$$

$$= \left(1+(f_2,f_1)/\|f_1\|^2\right)\|f_1\| \ge (1-\vartheta)\|f_1\|$$

Moreover, by Pythagoras' Theorem

$$\|f_1+f_2\|^2 \ge \|(f_1+f_2)_1^{\rm P}\|^2.$$

$\square$

It is then straightforward to derive the following result.

Corollary 3.1 *Suppose that $\vartheta(S,N) < 1/L$. Then*

$$\phi^2(L,S,N) \ge \left(1 - L\vartheta(S,N)\right)^2 \Lambda^2(S,N).$$

A similar result is true for the adaptive versions. In other words, the (adaptive) restricted regression condition implies the (adaptive) restricted eigenvalue condition.

4 S-coherence conditions imply adaptive (S,s)-restricted regression conditions

Bunea et al. (2007a,b,c) establish oracle results under a condition which we refer to as the restricted diagonal condition. They provide coherence conditions for verifying the restricted diagonal condition.

Definition: Restricted diagonal condition. *We say that the S-*restricted diagonal condition *holds if for some constant $\varphi(S) > 0$*

$$\Sigma - \varphi(S){\rm diag}(\iota_S)$$

is positive semi-definite. Here $\iota := (1,\ldots,1)^T$ (so $\iota_{j,S} = {\rm l}\{j \in S\}$).

We now show that coherence conditions actually imply restricted regression conditions. First, we consider some matrix norms in more detail. Let $1 \le q \le \infty$, and r be its conjugate, i.e.,

$$\frac{1}{q} + \frac{1}{r} = 1.$$

Define

$$\|\Sigma_{1,2}(\mathcal{N})\|_{2,q} := \sup_{\|\beta_{\mathcal{N}^c}\|_r \le 1} \|\Sigma_{1,2}(\mathcal{N})\beta_{\mathcal{N}^c}\|_2.$$

Some properties. *The quantity $\|\Sigma_{1,2}(\mathcal{N})\|_{2,2}^2$ is the largest eigenvalue of the matrix $\Sigma_{1,2}(\mathcal{N})\Sigma_{2,1}(\mathcal{N})$. We further have for $1 \le q < \infty$,*

$$\|\Sigma_{1,2}(\mathcal{N})\|_{2,q} \le \left(\sum_{j \notin \mathcal{N}} \left(\sqrt{\sum_{k \in \mathcal{N}} \sigma_{j,k}^2}\right)^q\right)^{1/q},$$

and similarly for $q = \infty$,

$$\|\Sigma_{1,2}(\mathcal{N})\|_{2,\infty} \le \max_{j \notin \mathcal{N}} \sqrt{\sum_{k \in \mathcal{N}} \sigma_{j,k}^2}.$$

Moreover,

$$\|\Sigma_{1,2}(\mathcal{N})\|_{2,q} \ge \|\Sigma_{1,2}(\mathcal{N})\|_{2,\infty},$$

so for replacing $\|\Sigma_{1,2}(\mathcal{N})\|_{2,\infty}$ by $\|\Sigma_{1,2}(\mathcal{N})\|_{2,q}$, $q < \infty$, one might have to pay a price.

Lemma 4.1 *For all $1 \le q \le \infty$, the following inequality holds:*

$$\vartheta_{\text{adaptive}}(S, 2s) \le \max_{\mathcal{N} \supset S,\ |\mathcal{N}|=2s} \frac{\sqrt{s}\|\Sigma_{1,2}(\mathcal{N})\|_{2,q}}{s^{1/q}\Lambda^2(S, 2s)}.$$

Moreover,

$$\vartheta_{\text{adaptive}}(S, s) \le \frac{\sqrt{s}\|\Sigma_{1,1}(S)\|_{2,\infty}}{\Lambda^2(S, s)}.$$

Proof of Lemma 4.1. Take r such that $1/q + 1/r = 1$. Let $\mathcal{N} \supset S$, with $|\mathcal{N}| = s$ and let $\beta \in \mathcal{R}_{\text{adaptive}}(S, \mathcal{N})$.

We let $f_{\mathcal{N}} := f_{\beta_{\mathcal{N}}}$, $f_{\mathcal{N}^c} := f_{\beta_{\mathcal{N}^c}}$.

We have

$$\begin{aligned} |(f_{\mathcal{N}}, f_{\mathcal{N}^c})| &= |\beta_{\mathcal{N}}^T \Sigma_{1,2}(\mathcal{N})\beta_{\mathcal{N}^c}| \\ &\le \|\Sigma_{1,2}(\mathcal{N})\|_{2,q}\|\beta_{\mathcal{N}^c}\|_r\|\beta_{\mathcal{N}}\|_2. \end{aligned}$$

Applying Lemma 2.2 gives

$$\|\beta_{\mathcal{N}^c}\|_r \le \|\beta_{S^c}\|_1/s^{1/q} \le \sqrt{s}\|\beta_S\|_2/s^{1/q} \le \sqrt{s}\|\beta_{\mathcal{N}}\|_2/s^{1/q}. \tag{3}$$

This yields

$$\begin{aligned} |(f_{\mathcal{N}}, f_{\mathcal{N}^c})| &\le \sqrt{s}\|\Sigma_{1,2}(S)\|_{2,q}\|\beta_{\mathcal{N}}\|_2^2/s^{1/q} \\ &\le \sqrt{s}\|\Sigma_{1,2}(S)\|_{2,q}\|f_{\mathcal{N}}\|_2^2/(s^{1/q}\Lambda^2(S, 2s)). \end{aligned}$$

Similarly,

$$\begin{aligned} |(f_S, f_{S^c})| &\le \|\Sigma_{1,2}(S)\|_{2,\infty}\|\beta_{S^c}\|_1\|\beta_S\|_2 \\ &\le \sqrt{s}\|\Sigma_{1,2}(S)\|_{2,\infty}\|\beta_S\|_2^2 \le \sqrt{s}\|\Sigma_{1,2}(S)\|_{2,\infty}/\Lambda^2(S, s). \end{aligned}$$

□

One of the consequences is in the spirit of the mutual coherence condition in Bunea et al. (2007b).

Corollary 4.1 *(Coherence with $q = \infty$) We have*

$$\vartheta_{\text{adaptive}}(S, s) \le \frac{\sqrt{s}\max_{j \notin S}\sqrt{\sum_{k \in S}\sigma_{j,k}^2}}{\Lambda^2(S, s)} \le \vartheta_{\text{mutual}}(S).$$

With $q = 1$ and $N = s$, the coherence lemma is similar to the cumulative local coherence condition in Bunea et al. (2007c). We also consider the case $N = 2s$.

Corollary 4.2 *(Coherence with $q = 1$) We have*

$$\vartheta_{\text{adaptive}}(S, s) \le \vartheta_{\text{cumulative}}(S),$$

and

$$\vartheta(S, 2s) \le \max_{\mathcal{N} \supset S,\ |\mathcal{N}|=2s} \frac{\sqrt{\sum_{k\in\mathcal{N}} \left(\sum_{j\notin\mathcal{N}} |\sigma_{j,k}| \right)^2 }}{\sqrt{s}\Lambda^2(S, 2s)}.$$

The coherence lemma with $q = 2$ is a condition about eigenvalues (recall that $\|\Sigma_{1,2}(\mathcal{N})\|_{2,2}^2$ equals the largest eigenvalue of $\Sigma_{1,2}(\mathcal{N})\Sigma_{2,1}(\mathcal{N})$). The bound is then much rougher than the one following from the weak $(S, 2s)$-restricted isometry condition, which we derive in Lemma 7.1.

Corollary 4.3 *(Coherence with $q = 2$) We have*

$$\vartheta_{\text{adaptive}}(S, 2s) \le \max_{\mathcal{N} \supset S,\ |\mathcal{N}|=2s} \frac{\|\Sigma_{1,2}(\mathcal{N})\|_{2,2}}{\Lambda^2(S, 2s)}.$$

5 The adaptive (S, s)-restricted regression condition implies the (S, s)-uniform irrepresentable condition

Theorem 5.1 *We have*

$$\vartheta_{\text{irrepresentable}}(S, s) \le \vartheta_{\text{adaptive}}(S, s).$$

Proof of Theorem 5.1. First observe that

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty = \sup_{\|\beta_{S^c}\|_1 \le 1} |\beta_{S^c}^T \Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S|$$

$$= \sup_{\|\beta_{S^c}\|_1 \le 1} |(f_{\beta_{S^c}}, f_{b_S})|,$$

where

$$b_S := \Sigma_{1,1}^{-1}(S)\tau_S.$$

We note that

$$\frac{\|f_{b_S}\|^2}{\sqrt{s}\|b_S\|_2} = \frac{\|\Sigma_{1,1}^{1/2}(S)b_S\|_2^2}{\|\Sigma_{1,1}(S)b_S\|_2\|b_S\|_2} \frac{\|\Sigma_{1,1}(S)b_S\|_2}{\sqrt{s}} \le 1.$$

(Use Cauchy-Schwarz inequality for bounding the first factor). Furthermore, for any constant c,

$$\sup_{\|\beta_{S^c}\|_1 \le 1} |(f_{\beta_{S^c}}, f_{b_S})| = \sup_{\|\beta_{S^c}\|_1 \le c} |(f_{\beta_{S^c}}, f_{b_S})|/c.$$

Take $c = \sqrt{s}\|\beta_S\|_2$ to find

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_\infty = \sup_{\|\beta_{S^c}\|_1 \le \sqrt{s}\|b_S\|_2} \frac{|(f_{\beta_{S^c}}, f_{b_S})|}{\sqrt{s}\|b_S\|_2}$$

$$\le \sup_{\|\beta_{S^c}\|_1 \le \sqrt{s}\|b_S\|_2} \frac{|(f_{\beta_{S^c}}, f_{b_S})|}{\|f_{b_S}\|^2}.$$

$\square$

6 The (S, s)-irrepresentable condition is sufficient and essentially necessary for variable selection

An important characterization of the solution β^* can be derived from the *Karush-Kuhn-Tucker* (*KKT*) conditions which in our context involves subdifferential calculus: see Bertsimas and Tsitsiklis (1997).

The KKT conditions. *We have*

$$2\Sigma(\beta^* - \beta^0) = -\lambda\tau^*.$$

Here $\|\tau^*\|_\infty \le 1$*, and moreover*

$$\tau_j^* \mathrm{l}\{\beta_j^* \ne 0\} = \mathrm{sign}(\beta_j^*),\ j = 1, \ldots, p.$$

For $\mathcal{N} \supset S$, we write the projection of a function f on the space spanned by $\{\psi_j\}_{j\in\mathcal{N}}$ as $f^{P_\mathcal{N}}$, and the anti-projection as $f^{A_\mathcal{N}} := f - f^{P_\mathcal{N}}$. Hence, we note that

$$f_\beta^{P_\mathcal{N}} = (f_{\beta_\mathcal{N}} + f_{\beta_{\mathcal{N}^c}})^{P_\mathcal{N}} = f_{\beta_\mathcal{N}} + (f_{\beta_{\mathcal{N}^c}})^{P_\mathcal{N}},$$

and thus

$$f_\beta^{A_\mathcal{N}} = (f_{\beta_{\mathcal{N}^c}})^{A_\mathcal{N}}.$$

Moreover

$$\|(f_{\beta_{\mathcal{N}^c}})^{A_\mathcal{N}}\|^2 = \beta_{\mathcal{N}^c}^T \Sigma_{2,2}(\mathcal{N})\beta_{\mathcal{N}^c} - \beta_{\mathcal{N}^c}^T \Sigma_{2,1}(\mathcal{N})\Sigma_{1,1}^{-1}(\mathcal{N})\Sigma_{1,2}(\mathcal{N})\beta_{\mathcal{N}^c}.$$

Lemma 6.1 *Suppose* $\Sigma_{1,1}^{-1}(\mathcal{N})$ *exists. We have*

$$2\|(f_{\beta^*_{\mathcal{N}^c}})^{A_\mathcal{N}}\|^2 = \lambda(\beta^*_{\mathcal{N}^c})^T \Sigma_{2,1}(\mathcal{N})\Sigma_{1,1}^{-1}(\mathcal{N})\tau^*_\mathcal{N} - \lambda\|\beta^*_{\mathcal{N}^c}\|_1.$$

Proof of Lemma 6.1. By the KKT conditions, we must have

$$2\Sigma_{1,1}(\mathcal{N})(\beta^*_\mathcal{N} - \beta^0_\mathcal{N}) + 2\Sigma_{1,2}(\mathcal{N})\beta^*_{\mathcal{N}^c} = -\lambda\tau^*_\mathcal{N},$$

$$2\Sigma_{2,1}(\mathcal{N})(\beta^*_\mathcal{N} - \beta^0_\mathcal{N}) + 2\Sigma_{2,2}(\mathcal{N})\beta^*_{\mathcal{N}^c} = -\lambda\tau^*_{\mathcal{N}^c}.$$

It follows that

$$2(\beta^*_{\mathcal{N}} - \beta^0_{\mathcal{N}}) + 2\Sigma^{-1}_{1,1}(\mathcal{N})\Sigma_{1,2}(\mathcal{N})\beta^*_{\mathcal{N}^c} = -\lambda\Sigma^{-1}_{1,1}(\mathcal{N})\tau^*_{\mathcal{N}},$$

$$2\Sigma_{2,1}(\mathcal{N})(\beta^*_{\mathcal{N}} - \beta^0_{\mathcal{N}}) + 2\Sigma_{2,2}(\mathcal{N})\beta^*_{\mathcal{N}^c} = -\lambda\tau^*_{\mathcal{N}^c}$$

(leaving the second equality untouched). Hence, multiplying the first equality by $-(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,1}(\mathcal{N})$, and the second by $(\beta^*_{\mathcal{N}^c})^T$,

$$-2(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,1}(\mathcal{N})(\beta^*_{\mathcal{N}} - \beta^0_{\mathcal{N}}) - 2(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,1}(\mathcal{N})\Sigma^{-1}_{1,1}(\mathcal{N})\Sigma_{1,2}(\mathcal{N})\beta^*_{\mathcal{N}^c}$$

$$= \lambda(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,1}(\mathcal{N})\Sigma^{-1}_{1,1}(\mathcal{N})\tau^*_{\mathcal{N}},$$

$$2(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,1}(\mathcal{N})(\beta^*_{\mathcal{N}} - \beta^0_{\mathcal{N}}) + 2(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,2}(\mathcal{N})\beta^*_{\mathcal{N}^c} = -\lambda\|\beta^*_{\mathcal{N}^c}\|_1,$$

where we invoked that $\beta^*_j\tau^*_j = |\beta^*_j|$. Adding up the two equalities gives

$$2(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,2}(\mathcal{N})\beta^*_{\mathcal{N}^c} - 2(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,1}(\mathcal{N})\Sigma^{-1}_{1,1}(\mathcal{N})\Sigma_{1,2}(\mathcal{N})\beta^*_{\mathcal{N}^c}$$

$$= \lambda(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,1}(\mathcal{N})\Sigma^{-1}_{1,1}(\mathcal{N})\tau^*_{\mathcal{N}} - \lambda\|\beta^*_{\mathcal{N}^c}\|_1.$$

□

We now connect the irrepresentable condition to variable selection. Define

$$|\beta^0|_{\min} := \min\{|\beta^0_j| : \ j \in S\}.$$

Lemma 6.2
Part 1. *Suppose the (S, N)-uniform irrepresentable condition holds. Then $|S_*\backslash S| \le N - s$.*
Part 2. *Suppose the (S, N)-irrepresentable condition holds and*

$$|\beta^0_{\min}| > \lambda s/\phi^2_{\text{compatible}}(S).$$

Then $S_* \supset S$ and $|S_*| \le N$.
Part 3. *Conversely, suppose that $S_* \supset S$ and $|S_*| \le N$, and $\Lambda(S, N) > 0$. Then*

$$\|\Sigma_{2,1}(S_*)\Sigma^{-1}_{1,1}(S_*)\tau^*_{S_*}\|_\infty \le 1.$$

If moreover

$$|\beta^0|_{\min} > \lambda\sqrt{s}/(2\Lambda(S, N)),$$

then $\tau^*_{S_*} = \tau^0_{S_*}$, where $\tau^0_{S_*} := \text{sign}(\beta^0_{S_*})$.

A special case is $N = s$. In Part 1, we then obtain that $S_* \subset S$, i.e., no false positive selections. Moreover, Part 2 then proves $S_* = S$ and Part 3 assumes $S_* = S$.

Proof of Lemma 6.2.

Part 1. Let $\mathcal{N} \supset S$ be a set of size at most N, such that

$$\sup_{\|\tau_S\|_\infty \le 1} \|\Sigma_{2,1}(\mathcal{N})\Sigma^{-1}_{1,1}(\mathcal{N})\tau_{\mathcal{N}}\|_\infty < 1.$$

By Lemma 6.1, we now have that if $\|\beta^*_{\mathcal{N}^c}\|_1 > 0$

$$2\|(f^*)^{A_{\mathcal{N}}}\|^2 = \lambda(\beta^*_{\mathcal{N}^c})^T\Sigma_{2,1}(\mathcal{N})\Sigma_{1,1}^{-1}(\mathcal{N})\tau^*_{\mathcal{N}} - \lambda\|\beta^*_{\mathcal{N}^c}\|_1 < 0,$$

which is a contradiction. Hence $\|\beta^*_{\mathcal{N}^c}\|_1 = 0$, i.e., $S_* \subset \mathcal{N}$.

Part 2. By Lemma 2.1,

$$\|\beta^*_S - \beta^0_S\|_1 \le \sqrt{s}\|f^* - f^0\|/\phi_{\rm compatible}(S) \le \lambda s/\phi^2_{\rm compatible}(S).$$

The condition $|\beta^0_{\min}| > \lambda s/\phi^2_{\rm compatible}(S)$ thus implies that $S_* \supset S$, and hence that $\tau^*_S \in \{-1,1\}^s$. We also know that $\tau^*_{S_*} \in \{-1,1\}$. Hence for any $\mathcal{N}$ satisfying $S \subset \mathcal{N} \subset S_*$, also $\tau_{\mathcal{N}} \in \{-1,1\}^{|\mathcal{N}|}$. Thus, by the (S,N)-irrepresentable condition, there exists such an $\mathcal{N}$, say $\tilde{\mathcal{N}}$, with

$$\|\Sigma_{2,1}(\tilde{\mathcal{N}})\Sigma^{-1}_{1,1}(\tilde{\mathcal{N}})\tau^*_{\tilde{\mathcal{N}}}\|_\infty < 1.$$

As in Part 1, we then must have that $\|\beta^*_{\tilde{\mathcal{N}}^c}\|_1 = 0$.

Part 3. Because $\Lambda(S,N) > 0$, and $|S_*| \le N$, we know that $\Sigma^{-1}_{1,1}(S_*)$ exists. Because $S_* \supset S$, we have $\beta^*_{S_*^c} = \beta^0_{S_*^c} = 0$, so the KKT conditions take the form

$$2\Sigma_{1,1}(S_*)(\beta^*_{S_*} - \beta^0_{S_*}) = -\lambda\tau^*_{S_*},$$

and

$$2\Sigma_{2,1}(S_*)(\beta^*_{S_*} - \beta^0_{S_*}) = -\lambda\tau^*_{S_*^c}.$$

Hence

$$\beta^*_{S_*} - \beta^0_{S_*} = \lambda\Sigma^{-1}_{1,1}(S_*)\tau^*_{S_*}/2,$$

and, inserting this in the second KKT equality,

$$\Sigma_{2,1}(S_*)\Sigma^{-1}_{1,1}(S_*)\tau^*_{S_*} = \tau^*_{S_*^c}.$$

But then

$$\|\Sigma_{2,1}(S_*)\Sigma^{-1}_{1,1}(S_*)\tau^*_{S_*}\|_\infty = \|\tau^*_{S_*^c}\|_\infty \le 1.$$

The first KKT equality moreover implies

$$\|\beta^*_{S_*} - \beta^0_{S_*}\|_2 \le \lambda\sqrt{N}/(2\Lambda^2(S,N)).$$

So when $|\beta^0|_{\min} > \lambda\sqrt{N}/(2\Lambda^2(S,N))$, we have $\tau^*_{S_*} = \tau^0_{S_*}$.

$\square$

7 The weak $(S,2s)$-restricted isometry property implies the $(S,2s)$-restricted regression condition

Lemma 7.1 *We have*

$$\vartheta_{\rm adaptive}(S,2s) \le \vartheta_{\rm weak-RIP}(S,2s).$$

Proof of Lemma 7.1. Let β be an arbitrary vector. satisfying $\|\beta_{S^c}\|_1 \le \sqrt{s}\|\beta_S\|_2$. From Lemma 2.2,

$$\sum_{k=1}^K \|\beta_{\mathcal{N}_k}\|_2 \le \|\beta_{S^c}\|_1/\sqrt{s} \le \|\beta_S\|_2.$$

Hence, using the definition of the restricted orthogonality constant $\theta(S, 2s)$, and of the $(S, 2s)$-uniform eigenvalue $\Lambda^2(S, 2s)$,

$$|(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{N}^c}})| \le \theta(S, 2s) \sum_{k=1}^K \|\beta_{\mathcal{N}}\|_2 \|\beta_{\mathcal{N}_k}\|_2 \le \theta(S, 2s)\|\beta_{\mathcal{N}}\|_2\|\beta_S\|_2$$

$$\le \theta(S, 2s)\|f_{\beta_{\mathcal{N}}}\|_2^2/\Lambda^2(S, 2s),$$

or

$$\frac{|(f_{\beta_{\mathcal{N}}}, f_{\beta_{\mathcal{N}^c}})|}{\|f_{\beta_{\mathcal{N}}}\|^2} \le \theta(S, 2s)/\Lambda^2(S, 2s) = \vartheta_{\text{weak}-\text{RIP}}(S, 2s).$$

□

Corollary 7.1 *Together with Corollary 3.1, we can now conclude that when* $\vartheta_{\text{weak}-\text{RIP}}(S, 2s) < 1/L$*, one has*

$$\phi(L, S, 2s) \ge (1 - L\vartheta_{\text{weak}-\text{RIP}})^2 \Lambda^2(S, 2s).$$

This result is from Koltchinskii (2009a) and Bickel et al. (2009).

8 The restricted isometry property with small constants implies the weak $(S, 2s)$-irrepresentable condition

We start with two preparatory lemmas. Recall that

$$\vartheta_{\text{weak}-\text{RIP}}(S, s) = \theta(S, s)/\Lambda^2(S, s).$$

Lemma 8.1 *Suppose that*

$$\vartheta_{\text{weak}-\text{RIP}}(S, s) < 1.$$

Then

$$2\|(f_{\beta^*_{S^c}})^{A_S}\|^2 \le \vartheta_{\text{weak}-\text{RIP}}(S, s)\left(\lambda\sqrt{s}\|\beta^*_{\mathcal{N}_0^*}\|_2\right),$$

where A_S denotes the anti-projection defined in Section 6.

Proof of Lemma 8.1. Define

$$b_S := \Sigma_{1,1}(S)^{-1}\tau_S^*.$$

Then

$$\|b_S\| \le \|\tau_S^*\|_2/\Lambda^2(S, s) \le \sqrt{s}/\Lambda^2(S, s).$$

Moreover,

$$|(\beta_{S^c}^*)^T \Sigma_{2,1}\Sigma_{1,1}^{-1}(S)\tau_S^*| = |(f_{\beta_{S^c}^*}, f_{b_S})| \le \sum_{k=0}^{K-1} |(f_{\beta_{\mathcal{N}_k^*}^*}, f_{b_S})|$$

$$\le \theta(S,s)\sum_{k=0}^{K} \|\beta_{\mathcal{N}_k^*}^*\|_2 \|b_S\|_2 \le \theta(S,s)\|b_S\|_2 \left(\|\beta_{\mathcal{N}_0^*}^*\|_2 + \sum_{k=1}^{K} \|\beta_{\mathcal{N}_k^*}^*\|_2 \right)$$

$$\le \theta(S,s)\|b_S\|_2 \left(\|\beta_{\mathcal{N}_0^*}^*\|_2 + \|\beta_{S^c}^*\|_1/\sqrt{s} \right) \le \frac{\theta(S,s)}{\Lambda^2(S,s)}\sqrt{s}\|\beta_{\mathcal{N}_0^*}^*\|_2 + \frac{\theta(S,s)}{\Lambda^2(S,s)}\|\beta_{S^c}^*\|_1$$

$$= \vartheta_{\mathrm{weak-RIP}}(S,s)\biggl(\sqrt{s}\|\beta_{\mathcal{N}_0^*}^*\|_2 + \|\beta_{S^c}^*\|_1 \biggr).$$

Thus,

$$(\beta_{S^c}^*)^T \Sigma_{2,1}\Sigma_{1,1}^{-1}(S)\tau_S^* - \|\beta_{S^c}^*\|_1$$

$$\le \vartheta_{\mathrm{weak-RIP}}(S,s)\sqrt{s}\|\beta_{\mathcal{N}_0^*}^*\|_2 - (1-\vartheta_{\mathrm{weak-RIP}}(S,s))\|\beta_{S^c}^*\|_1$$

$$\le \vartheta_{\mathrm{weak-RIP}}(S,s)\sqrt{s}\|\beta_{\mathcal{N}_0^*}^*\|_2.$$

Hence, by Lemma 6.1,

$$2\|(f_{\beta_{S^c}^*})^{A_S}\|^2 \le \vartheta_{\mathrm{weak-RIP}}(S,s)\biggl(\lambda\sqrt{s}\|\beta_{\mathcal{N}_0^*}^*\|_2 \biggr).$$

□

Lemma 8.2 *Suppose that*

$$\vartheta_{\mathrm{weak-RIP}}(S,s) < 1.$$

Then for any subset $\tilde{\mathcal{N}} \subset S^c$, *with* $|\tilde{\mathcal{N}}| \le s$, *and any* $b \in \mathbb{R}^p$

$$|(f_{b_{\tilde{\mathcal{N}}}}, f^* - f^0)| \le \frac{\lambda\sqrt{s}}{\phi(S,2s)\Lambda(S,s)} \left(\theta(S,s) + \sqrt{(1+\delta_{s,s})\theta(S,s)/2} \right) \|b_{\tilde{\mathcal{N}}}\|_2.$$

Proof of Lemma 8.2. We have

$$|(f_{b_{\tilde{\mathcal{N}}}}, f^* - f^0)| \le |(f_{b_{\tilde{\mathcal{N}}}}, (f^* - f^0)^{P_S})| + |(f_{b_{\tilde{\mathcal{N}}}}, (f^*)^{A_S})|$$

Let us write

$$(f^* - f^0)^{P_S} := f_{\gamma_S}.$$

Then, invoking Lemma 2.1,

$$\|\gamma_S\|_2 \le \|f_{\gamma_S}\|/\Lambda(S,s) = \|(f^* - f^0)^{P_S}\|/\Lambda(S,s) \le \|f^* - f^0\|/\Lambda(S,s)$$

$$\le \lambda\sqrt{s}/\biggl(\phi(S,2s)\Lambda(S,s) \biggr).$$

It follows that

$$|(f_{b_{\tilde{\mathcal{N}}}}, (f^* - f^0)^{P_S})| \leq \theta(S,s)\|b_{\tilde{\mathcal{N}}}\|_2\|\gamma_S\|_2$$

$$\leq \theta(S,s)\|b_{\tilde{\mathcal{N}}}\|_2\lambda\sqrt{s}/\bigg(\phi(S,2s)\Lambda(S,s)\bigg).$$

Moreover, we have

$$\|\beta^*_{\mathcal{N}_0^*}\|_2 \leq \|\beta^*_{\mathcal{N}_*} - \beta^0_{\mathcal{N}_*}\|_2 \leq \lambda\sqrt{s}/\phi^2(S,2s).$$

So, by Lemma 8.1,

$$\|(f_{\beta^*_{S^c}})^{A_S}\|^2 \leq \frac{\theta(S,s)}{\Lambda^2(S,s)}\lambda\sqrt{s}\|\beta^*_{\mathcal{N}_*} - \beta^0_{\mathcal{N}_*}\|_2/2$$

$$\leq \lambda^2 s\theta(S,s)/\bigg(2\phi^2(S,2s)\Lambda^2(S,s)\bigg).$$

Therefore

$$|(f_{b_{\tilde{\mathcal{N}}}}, (f^*)^{A_S})| \leq \|f_{b_{\tilde{\mathcal{N}}}}\|\|(f^*)^{A_S})\| \leq \lambda\sqrt{s}\sqrt{\theta(S,s)/2}/\bigg(\phi(S,2s)\Lambda(S,s)\bigg)\|f_{b_{\tilde{\mathcal{N}}}}\|$$

$$\leq \frac{\sqrt{(1+\delta_s)\theta(S,s)/2}}{\phi(S,2s)\Lambda(S,s)}\lambda\sqrt{s}\|b_{\tilde{\mathcal{N}}}\|_2.$$

The next result shows that if the constants are small enough, then there will be no more than s false positives. We define

$$\alpha(S) := \frac{\bigg(\sqrt{2}\theta(S,s) + \sqrt{(1+\delta_s)\theta(S,s)}\bigg)}{\phi(S,2s)\Lambda(S,s)}. \tag{4}$$

Lemma 8.3 *Suppose that*

$$\alpha(S) < 1.$$

Then $|S_*\backslash S| < s$.

Proof of Lemma 8.3 Since $\alpha(S) < 1$, Lemma 8.2 implies that for any $\tilde{\mathcal{N}} \subset S^c$, with $|\tilde{\mathcal{N}}| \leq s$, and for any b with $\|b_{\tilde{\mathcal{N}}}\|_2 \neq 0$,

$$|(f_{b_{\tilde{\mathcal{N}}}}, f^* - f_0)| < \lambda\sqrt{s/2}\|b_{\tilde{\mathcal{N}}}\|_2.$$

Hence, taking $b_j = (\psi_j, f^* - f^0)$, $j \in \tilde{\mathcal{N}}$,

$$\sum_{j\in\tilde{\mathcal{N}}} |(\psi_j, f^* - f^0)|^2 < \lambda^2 s/2.$$

For $j \in S_*\backslash S$ we have by the KKT conditions

$$|2(\psi_j, f^* - f^0)| \geq \lambda.$$

Suppose now that $|S_* \backslash S| \geq s$. Then there is a subset $\mathcal{N}'$ of $S_* \backslash S$, with size $|\mathcal{N}'| = s$, and we have

$$\lambda^2 s/2 > \sum_{j \in \mathcal{N}'} |(\psi_j, f^* - f^0)|^2 \geq \lambda^2 |\mathcal{N}'|/2.$$

This is a contraction, and hence $|S_* \backslash S| < s$. □

This leads to the following result.

Theorem 8.1 *Suppose that $\alpha(S) < 1$, see (4). Then the weak $(S, 2s)$-irrepresentable condition holds.*

Proof of Theorem 8.1. As $\alpha(S) < 1$, we know that $\phi(S, 2s) > 0$. Take an arbitrary $\tau_S^0 \in \{-1, 1\}^s$, and a β_0 satisfying $\beta_S^0 = \beta^0$, $\text{sign}(\beta_S^0) = \tau_S^0$, and

$$|\beta^0|_{\min} > \lambda\sqrt{s}/\phi^2(S, 2s).$$

By Lemma 2.1, the Lasso satisfies

$$\|\beta_S^* - \beta_S^0\|_2 \leq \lambda\sqrt{s}/\phi^2(S, 2s).$$

Hence, we must have $S_* \supset S$, and $\tau_S^* = \tau_S^0$. Moreover, by Lemma 8.3, $|S_*| < 2s$. By Part 3 of Lemma 6.2, we must have

$$\|\Sigma_{2,1}(S_*)\Sigma_{1,1}^{-1}(S_*)\tau_{S_*}^*\|_\infty \leq 1.$$

Since $\tau_S^0 = \tau_S^*$ is arbitrary and $\tau_{S_*}^* \in \{-1, 1\}^{|S_*|}$, we conclude that the weak $(S, 2s)$-irrepresentable condition holds (in fact the weak $(S, 2s-1)$-irrepresentable condition holds).

□

Corollary 8.1 *The RIP is the condition $\vartheta_{\text{RIP}} < 1$, or equivalently*

$$\delta_s + \theta_{s,s} + \theta_{s,2s} < 1.$$

Candès and Tao (2005) show that $\delta_{2s} \leq \theta_s + \delta_s$. The restricted isometry constant δ_s has to be less than one, so we may use the bound $1 + \delta_s \leq 2$. Moreover, it is clear that $\theta(S, N) \leq \theta_{s,N}$, and $\Lambda^2(S, N) \geq 1 - \delta_N$. Inserting these bounds in Corollary 7.1 we find

$$\phi(S, 2s)\Lambda(S, s) \geq (1 - \delta_s - \theta_{s,s} - \theta_{s,2s})\sqrt{\frac{1 - \delta_s}{1 - \delta_s - \theta_{s,s}}} \geq (1 - \delta_s - \theta_{s,s} - \theta_{s,2s}).$$

It follows that

$$\alpha(S) \leq \frac{\sqrt{2}(\theta_{s,s} + \sqrt{\theta_{s,s}})}{1 - \delta_s - \theta_{s,s} - \theta_{s,2s}}.$$

For example, if $\delta_s \leq \sqrt{2} - 1$ and $\theta_{s,2s} \leq \frac{1}{16}$, we get (invoking $\theta_{s,s} \leq \theta_{s,2s}$)

$$\alpha(S) \leq 0.96.$$

We conclude that the RIP with small enough constants implies the weak $(S, 2s)$-irrepresentable condition.

As Candès and Tao (2005) show, the RIP implies exact recovery. To complete the picture, we now show that the (S, s)-irrepresentable condition also implies exact recovery.

The linear programming problem is

$$\min\{\|\beta\|_1 : \ \|f_\beta - f^0\| = 0\},$$

where, as before $f^0 = f_{\beta^0}$ with $\beta^0 = \beta^0_S$. Let β^{LP} be the minimizer of the linear programming problem.

Lemma 8.4 *Suppose the (S, s)-irrepresentable condition holds. Then one has exact recovery, i.e., $\beta^{\mathrm{LP}} = \beta^0$.*

Proof of Lemma 8.4. This follows from Candès and Tao (2005). They show that $\beta^{\mathrm{LP}} = \beta^0$ if one can find a $g \in L_2(P)$, such that
(i) $(\psi_j, g) = \tau^0_j$, for all $j \in S$,
(ii) $|(\psi_j, g)| < 1$ for all $j \notin S$,
where, as before, $\tau^0_S := \mathrm{sign}(\beta^0_S)$. The (S, s)-irrepresentable condition says that this is true for $g = f_{b_S}$, where $b_S = \Sigma^{-1}_{1,1}(S)\tau^0_S$. □

9 The (S, s)-uniform irrepresentable condition implies the S-compatibility condition

As the (S, s)-irrepresentable condition implies variable selection, one expects it will be more restrictive than the compatibility condition, which only implies a bound for the prediction error (and ℓ_1-estimation error). This turns out to be indeed the case, albeit we prove it only under the uniform version of the irrepresentable condition.

Theorem 9.1 *Suppose that*

$$\vartheta_{\mathrm{irrepresentable}}(S, s) < 1/L.$$

Then

$$\phi^2_{\mathrm{compatible}}(L, S) \geq (1 - L\vartheta_{\mathrm{irrepresentable}}(S, s))^2 \Lambda^2(S, s).$$

Proof of Theorem 9.1. Define,

$$\beta^\diamond := \arg\min_\beta \{\|f_\beta\|^2 : \ \|\beta_S\|_1 = 1, \|\beta_{S^c}\|_1 \leq L\}.$$

Let us write $f^\diamond := f^\diamond_\beta$, $f^\diamond_S := f_{\beta^\diamond_S}$ and $f^\diamond_{S^c} := f_{\beta^\diamond_{S^c}}$. Introduce a Lagrange multiplier $\lambda \in \mathbb{R}$ for the constraint $\|\beta_s\|_1 = 1$. By the KKT conditions, there exists a vector $\tau^\diamond_S$, with $\|\tau^\diamond_S\|_\infty \leq 1$, such that $\tau^T_S \beta^\diamond_S = \|\beta^\diamond_S\|_1$, and such that

$$\Sigma_{1,1}(S)\beta^\diamond_S + \Sigma_{1,2}(S)\beta^\diamond_{S^c} = -\lambda\tau^\diamond_S. \tag{5}$$

By multiplying by $(\beta^\diamond_S)^T$, we obtain

$$\|f^\diamond_S\|^2 + (f^\diamond_S, f^\diamond_{S^c}) = -\lambda\|\beta^\diamond_S\|_1.$$

The restriction $\|\beta_S^\diamond\|_1 = 1$ gives

$$\|f_S^\diamond\|^2 + (f_S^\diamond, f_{S^c}^\diamond) = -\lambda.$$

We also have from (5)

$$\beta_S^\diamond + \Sigma_{1,1}^{-1}(S)\Sigma_{1,2}(S)\beta_{S^c}^\diamond = -\lambda\Sigma_{1,1}^{-1}\tau_S^\diamond. \tag{6}$$

Hence, by multiplying with $(\tau_S^\diamond)^T$,

$$\|\beta_S^\diamond\|_1 + (\tau_S^\diamond)^T\Sigma_{1,1}^{-1}(S)\Sigma_{1,2}(S)\beta_{S^c}^\diamond = -\lambda(\tau_S^\diamond)^T\Sigma_{1,1}^{-1}\tau_S^\diamond,$$

or

$$\begin{aligned} 1 &= -(\tau_S^\diamond)^T\Sigma_{1,1}^{-1}(S)\Sigma_{1,2}(S)\beta_{S^c}^\diamond - \lambda(\tau_S^\diamond)^T\Sigma_{1,1}^{-1}(S)\tau_S^\diamond \\ &\le \vartheta\|\beta_{S^c}^\diamond\|_1 - \lambda(\tau_S^\diamond)^T\Sigma_{1,1}^{-1}(S)\tau_S^\diamond \\ &\le L\vartheta - \lambda(\tau_S^\diamond)^T\Sigma_{1,1}^{-1}(S)\tau_S^\diamond. \end{aligned}$$

Here, we applied that the (S, s)-uniform irrepresentable condition, with $\vartheta = \vartheta_{\text{irrepresentable}}(S, s)$, and the condition $\|\beta_{S^c}\|_1 \le L$. Thus

$$1 - L\vartheta \le -\lambda(\tau_S^\diamond)^T\Sigma_{1,1}^{-1}(S)\tau_S^\diamond.$$

Because $1 - L\vartheta > 0$ and $(\tau_S^\diamond)^T\Sigma_{1,1}^{-1}(S)\tau_S^\diamond \ge 0$, this implies that $\lambda < 0$, and in fact that

$$(1 - L\vartheta) \le -\lambda s/\Lambda^2(S, s),$$

where we invoked

$$(\tau_S^\diamond)^T\Sigma_{1,1}^{-1}(S)\tau_S^\diamond \le \|\tau_S^\diamond\|_2^2/\Lambda^2(S, s) \le s/\Lambda^2(S, s).$$

So

$$-\lambda \ge (1 - L\vartheta)\Lambda^2(S, s)/s.$$

Continuing with (6), we moreover have

$$\begin{aligned} &(\beta_{S^c}^\diamond)^T\Sigma_{2,1}(S)\beta_S^\diamond + (\beta_{S^c}^\diamond)^T\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\Sigma_{1,2}(S)\beta_{S^c}^\diamond \\ &= -\lambda(\beta_{S^c}^\diamond)^T\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S^\diamond. \end{aligned}$$

In other words,

$$(f_S^\diamond, f_{S^c}^\diamond) + \|(f_{S^c}^\diamond)^{P_S}\|^2 = -\lambda(\beta_{S^c}^\diamond)^T\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S^\diamond,$$

where $(f_{S^c}^\diamond)^{P_S}$ is the projection of $f_{S^c}^\diamond$ on the space spanned by $\{\psi_k\}_{k\in S}$. Again, by the (S, s)-uniform irrepresentable condition and by $\|\beta_{S^c}^\diamond\|_1 \le L$,

$$\left|(\beta_{S^c}^\diamond)^T\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S^\diamond\right| \le \vartheta\|\beta_{S^c}^\diamond\|_1 \le L\vartheta,$$

so

$$-\lambda(\beta_{S^c}^\diamond)^T\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S^\diamond = |\lambda|(\beta_{S^c}^\diamond)^T\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S^\diamond$$

$$\geq -|\lambda| \left| (\beta^{\diamond}_{S^c})^T \Sigma_{2,1}(S) \Sigma^{-1}_{1,1}(S) \tau^{\diamond}_S \right| \geq -|\lambda| L\vartheta = \lambda L\vartheta.$$

It follows that

$$\begin{aligned} \|f^{\diamond}\|^2 &= \|f^{\diamond}_S\|^2 + 2(f^{\diamond}_S, f^{\diamond}_{S^c}) + \|f^{\diamond}_{S^c}\|^2 \\ &= -\lambda + (f^{\diamond}_S, f^{\diamond}_{S^c}) + \|f^{\diamond}_{S^c}\|^2 \\ &\geq -\lambda + (f^{\diamond}_S, f^{\diamond}_{S^c}) + \|(f^{\diamond}_{S^c})^{\mathrm{P}}_S\|^2 \geq -\lambda + \lambda L\vartheta = -\lambda(1 - L\vartheta) \\ &\geq (1 - L\vartheta)^2 \Lambda^2(S, s)/s. \end{aligned}$$

Finally note that $\|f^{\diamond}\|^2 = \phi^2_{\text{compatible}}(L, S)/s$. □

10 Verifying the compatibility and restricted eigenvalue condition

In this section, we discuss the theoretical verification of the conditions. Determining a restricted ℓ_1-eigenvalue is in itself again a Lasso type of problem. Therefore, it is very useful to look for some good lower bounds.

A first, rather trivial, observation is that if Σ is non-singular, the restricted eigenvalue condition holds for all L, S and N, with $\phi^2(L, S, N) \geq \Lambda^2_{\min}(\Sigma)$, the latter being the smallest eigenvalue of Σ. If Σ is the population covariance matrix of a random design, i.e., the probability measure Q is the theoretical distribution of observed co-variables in $\mathcal{X}$, assuming positive definiteness of Σ is not very restrictive. We will present some examples in Section 10.2. Compatibility conditions for the population Gram matrix are of direct relevance if one replaces L_2-loss by robust convex loss (van de Geer, 2008). But, as we will show in the next subsection, even if Σ corresponds to the empirical covariance matrix of a fixed design, i.e., the measure Q is the empirical measure Q_n of n observed co-variables in $\mathcal{X}$, the compatibility and restricted eigenvalue condition is often "inherited" from the population version. Therefore, even for fixed designs (and singular Σ), the collection of cases where compatibility or restricted eigenvalue conditions hold is quite large.

10.1 Approximating the Gram matrix

For two (positive semi-definite) matrices Σ_0 and Σ_1, we define the supremum distance

$$d_\infty(\Sigma_1, \Sigma_0) := \max_{j,k} |(\Sigma_1)_{j,k} - (\Sigma_0)_{j,k}|.$$

Generally, perturbing the entries in Σ by a small amount may have a large impact on the eigenvalues of Σ. This is not true for (adaptive) restricted ℓ_1-eigenvalues, as is shown in the next lemma and its corollary.

Lemma 10.1 *Assume*

$$d_\infty(\Sigma_1, \Sigma_0) \leq \tilde{\lambda}.$$

Then $\forall\ \beta \in \mathcal{R}(L,S)$,

$$\left| \frac{\|f_\beta\|_{\Sigma_1}^2}{\|f_\beta\|_{\Sigma_0}^2} - 1 \right| \le \frac{(L+1)^2 \tilde{\lambda} s}{\phi_{\text{compatible}}^2(\Sigma_0, L, S)},$$

and similarly, $\forall\ \mathcal{N} \supset S,\ |\mathcal{N}| = N,$ *and* $\forall\ \beta \in \mathcal{R}(L,S,\mathcal{N})$,

$$\left| \frac{\|f_\beta\|_{\Sigma_1}^2}{\|f_\beta\|_{\Sigma_0}^2} - 1 \right| \le \frac{(L+1)^2 \tilde{\lambda} s}{\phi^2(\Sigma_0, L, S, N)},$$

and $\forall\ \mathcal{N} \supset S,\ |\mathcal{N}| = N,$ *and* $\forall\ \beta \in \mathcal{R}_{\text{adaptive}}(L,S,\mathcal{N})$,

$$\left| \frac{\|f_\beta\|_{\Sigma_1}^2}{\|f_\beta\|_{\Sigma_0}^2} - 1 \right| \le \frac{(L+1)^2 \tilde{\lambda} s}{\phi_{\text{adaptive}}^2(\Sigma_0, L, S, N)}.$$

Proof of Lemma 10.1. For all β,

$$\begin{aligned} \left| \|f_\beta\|_{\Sigma_1}^2 - \|f_\beta\|_{\Sigma_0}^2 \right| &= |\beta^T \Sigma_1 \beta - \beta^T \Sigma_0 \beta| \\ &= |\beta^T (\Sigma_1 - \Sigma_0) \beta| \le \tilde{\lambda} \|\beta\|_1^2. \end{aligned}$$

But if $\beta \in \mathcal{R}(L,S)$, it holds that $\|\beta_{S^c}\|_1 \le L \|\beta_S\|_1$, and hence

$$\|\beta\|_1 \le (L+1)\|\beta_S\|_1 \le (L+1)\|f_\beta\|_{\Sigma_0}\sqrt{s}/\phi_{\text{compatible}}(\Sigma_0, L, S).$$

This gives

$$\left| \|f_\beta\|_{\Sigma_1}^2 - \|f_\beta\|_{\Sigma_0}^2 \right| \le (L+1)^2 \tilde{\lambda} \|f_\beta\|_{\Sigma_0}^2 s / \phi_{\text{compatible}}^2(\Sigma_0, L, S).$$

The second result can be shown in the same way, and the third result as well as for $\beta \in \mathcal{R}_{\text{adaptive}}(L,S,\mathcal{N})$, it holds that $\|\beta_{S^c}\|_1 \le L\sqrt{s}\|\beta_S\|_2$, and hence

$$\|\beta\|_1 \le L\sqrt{s}\|\beta_S\|_2 + \|\beta_S\|_1 \le (L+1)\sqrt{s}\|\beta_S\|_2.$$

□

Corollary 10.1 *We have*

$$\phi_{\text{compatible}}(\Sigma_1, L, S) \ge \phi_{\text{compatible}}(\Sigma_0, L, S) - (L+1)\sqrt{d_\infty(\Sigma_0, \Sigma_1) s}.$$

Similarly,

$$\phi(\Sigma_1, L, S, N) \ge \phi(\Sigma_0, L, S, N) - (L+1)\sqrt{d_\infty(\Sigma_0, \Sigma_1) s},$$

and the same result holds for the adaptive version.

Corollary 10.1 shows that if one can find a matrix Σ_0 with well-behaved smallest eigenvalue, in a small enough ℓ_∞-neighborhood of Σ_1, then the restricted eigenvalue condition holds for Σ_1. As an example, consider the situation where $\psi_j(x) = x_j$ $(j = 1, \ldots, p)$ and where

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X}/n = (\hat{\sigma}_{j,k}),$$

where $\mathbf{X} = (X_{i,j})$ is a $(n \times p)$-matrix whose columns consist of i.i.d. $\mathcal{N}(0,1)$-distributed entries (but allowing for dependence between columns). We denote by Σ the population

covariance matrix of a row of $\mathbf{X}$. Using a union bound, it is not difficult to show that for all $t > 0$, and for

$$\tilde{\lambda}(t) := \sqrt{\frac{4t + 8\log p}{n}} + \frac{4t + 8\log p}{n},$$

one has the inequality

$$\mathbb{P}\left(d_\infty(\hat{\Sigma}, \Sigma) \geq \tilde{\lambda}(t) \right) \leq 2\exp[-t]. \tag{7}$$

This implies that if the smallest eigenvalue $\Lambda^2_{\min}(\Sigma)$ of Σ is bounded away from zero, and if the sparsity s is of smaller order $o(\sqrt{n/\log p})$, then the restricted eigenvalue condition holds with constant $\phi(S, N)$ not much smaller than $\Lambda_{\min}(\Sigma)$. The result can be extended to distributions with Gaussian tails.

10.2 Some examples

In the following, our discussion mainly applies for Σ being the population covariance matrix. For Σ being the empirical covariance matrix, the assumptions in the discussion below are unrealistic, but as seen in the previous section, the population properties can have important implications for the restricted eigenvalues of the empirical covariance matrix.

Example 10.1 *Consider the matrix*

$$\Sigma := (1 - \rho)I + \rho \iota \iota^T,$$

with $0 < \rho < 1$, *and* $\iota := (1, \ldots, 1)^T$ *a vector of 1's. Then the smallest eigenvalue of* Σ *is* $\Lambda^2_{\min}(\Sigma) = 1 - \rho$, *so the* (L, S, N)*-restricted eigenvalue condition holds with* $\phi^2(L, S, N) \geq 1 - \rho$. *The uniform* (S, s)*-irrepresentable condition is always met. The largest eigenvalue of* Σ *is* $(1 - \rho) + \rho p$. *Hence, the restricted isometry constants* δ_s *are defined only for* $\rho < 1/(s - 1)$.

Example 10.2 *In this example,* Σ *is a Toeplitz matrix, defined as follows. Consider a positive definite function*

$$R(k), \ k \in \mathbb{Z},$$

which is symmetric ($R(k) = R(-k)$) *and sufficiently regular in the following sense. The corresponding spectral density*

$$f_{\text{spec}}(\gamma) := \sum_{k=-\infty}^{\infty} R(k)\exp(-ik\gamma) \ (\gamma \in [-\pi, \pi])$$

is assumed to exist, to be continuous and periodic, and

$$\gamma_0 := \arg\min_{\gamma \in [0, \pi]} f_{\text{spec}}(\gamma)$$

is assumed unique, with $f(\gamma_0) = M > 0$. *Moreover, we suppose that* $f_{\text{spec}}(\cdot)$ *is* (2α) *continuously differentiable at* γ_0, *with* $f^{(2\alpha)}(\gamma_0) > 0$. *A Toeplitz matrix is*

$$\Sigma = (\sigma_{j,k}), \ \sigma_{j,k} := R(|j - k|),$$

where $R(\cdot)$ satisfies the conditions described above (in terms of the spectral density). A special case arises with $\sigma_{j,k} = \rho^{|j-k|}$ for some $0 \le \rho < 1$. The smallest eigenvalue $\Lambda^2_{\min}(\Sigma)$ of Σ is bounded away from zero where the bound is independent of p (Parter, 1961).

Example 10.3 *Consider a matrix Σ which is of block structure form:*

$$\Sigma = \ \mathrm{diag}(\Sigma_1, \ldots, \Sigma_k),$$

where the Σ_j are $(m \times m)$ covariance matrices ($j = 1, \ldots, k$) (the restriction to having the same dimension m can be easily dropped) and $km = p$. If the minimal eigenvalues satisfy

$$\min_j \Lambda^2_{\min}(\Sigma_j) \ge \eta^2 > 0,$$

then the minimal eigenvalue of Σ is also bounded from below by $\eta^2 > 0$. When m is much smaller than p, it is (much) less restrictive that small $m \times m$ covariance matrices Σ_j have well-behaved minimal eigenvalues than large $p \times p$ matrices.

Example 10.4 *This example presents a case where the compatibility condition holds, but where the uniform irrepresentable constant is very large. We also calculate the adaptive restricted regression. Let the first s indices $\{1, \ldots, s\}$ be the active set S and suppose that*

$$\Sigma := \begin{pmatrix} I & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix},$$

where I is the $(s \times s)$-identity matrix, and

$$\Sigma_{2,1} := \rho(b_2 b_1^T),$$

with $0 \le \rho < 1$, and with b_1 an s-vector and b_2 a $(p-s)$-vector, satisfying $\|b_1\|_2 = \|b_2\|_2 = 1$. Moreover, $\Sigma_{2,2}$ is some $(p-s) \times (p-s)$-matrix, with $\mathrm{diag}(\Sigma_{2,2}) = I$, and with smallest eigenvalue $\Lambda^2_{\min}(\Sigma_{2,2})$. One easily verifies that

$$\Lambda^2_{\min}(\Sigma) \ge \Lambda^2_{\min}(\Sigma_{2,2}) - \rho.$$

Moreover, for $b_1 := (1, 1, \ldots, 1)^T/\sqrt{s}$ and $b_2 := (1, 0, \ldots, 0)^T$, and $\rho > 1/\sqrt{s}$, the (S, s)-uniform irrepresentable condition does not hold, as in that case

$$\sup_{\|\tau_S\|_\infty \le 1} \|\Sigma_{2,1}(S)\Sigma^{-1}_{1,1}(S)\tau_S\|_\infty = \rho\sqrt{s}.$$

However, for any $N > s$, the (S, N)-uniform irrepresentable condition does hold. We moreover have

$$\vartheta_{\mathrm{adaptive}}(S) = \sqrt{s}\|\Sigma_{1,2}\|_{2,\infty} = \sqrt{s}\rho,$$

i.e. (since $\Lambda(S, s) = 1$), the bounds of Lemma 4.1 and Theorem 5.1 are strict in this example.

Example 10.5 *We recall that $\phi_{\mathrm{compatible}}(S) \ge \phi(S, s)$. Here is an example where the compatibility condition holds with reasonable $\phi^2_{\mathrm{compatible}}(S)$, but where the restricted eigenvalue $\phi^2(S, s)$ is very small. Assume $s > 2$. Let the first s indices $\{1, \ldots, s\}$ be the active set S with corresponding $(s \times s)$ covariance matrix $\Sigma_{1,1}$, and suppose that*

$$\Sigma := \mathrm{diag}(\Sigma_{1,1}, I),$$

where

$$\Sigma_{1,1} = \mathrm{diag}(B, I),$$

and, for some $0 \le \rho < 1 - 1/(s-2)$,

$$B = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We then have

$$\beta_S^T \Sigma_{1,1} \beta_S = (1-\rho)(\beta_1^2 + \beta_2^2) + \rho(\beta_1 + \beta_2)^2 + \sum_{j=3}^s \beta_j^2$$

$$\ge (1-\rho)(\beta_1^2 + \beta_2^2) + (\sum_{j=3}^s |\beta_j|)^2/(s-2)$$

Hence,

$$\min_{\|\beta_S\|_1=1} \beta_S^T \Sigma_{1,1} \beta_S \ge \min_{|\beta_1|+|\beta_2|\le 1} \left\{ (1-\rho)(\beta_1^2+\beta_2^2) + (1 - |\beta_1| - |\beta_2|)^2/(s-2) \right\}$$

$$\ge \min_{|\beta_1|+|\beta_2|\le 1} \left\{ \sum_{j=1,2} \left(1 - \rho + \frac{1}{s-2} \right) \beta_j^2 + \frac{1}{2-s} - 2\frac{|\beta_1| + |\beta_2|}{s-2} \right\}$$

$$= \min_{|\beta_1|+|\beta_2|\le 1} \left\{ \frac{(s-2)(1-\rho)+1}{s-2} \sum_{j=1,2} \left(|\beta_j| - \frac{1}{(s-2)(1-\rho)+1} \right)^2 \right\}$$

$$- \frac{2}{(s-2)\Big((s-2)(1-\rho)+1\Big)} + \frac{1}{s-2}$$

$$\ge \frac{(s-2)(1-\rho) - 1}{(s-2)\Big((s-2)(1-\rho)+1\Big)}.$$

It follows that

$$\phi^2_{\mathrm{compatible}}(S) = \min_{\|\beta_S\|_1=1,\ \|\beta_{S^c}\|_1 \le 1} \frac{s\beta^T \Sigma \beta}{\|\beta_S\|_1^2} \ge \frac{s\Big((s-2)(1-\rho)-1\Big)}{(s-2)\Big((s-2)(1-\rho)+1\Big)}$$

$$\ge \frac{(s-2)(1-\rho)-1}{(s-2)(1-\rho)+1}.$$

On the other hand

$$\phi^2(S,s) = \Lambda^2(S,s) = (1-\rho).$$

Hence, for example when $1-\rho = 3/(s-2)$, *we get*

$$\phi^2_{\mathrm{compatible}}(S) \ge 1/2$$

and

$$\phi^2(S, s) = \frac{3}{s-2}.$$

Clearly, for large s, this means that $\phi_{\text{compatible}}(S)$ is much better behaved than $\phi(S, s)$. Note that large s in this example (with $1 - \rho = 3/(s-2)$) corresponds to a correlation ρ close to one, i.e., to a case where Σ is "almost" singular.

11 Adding noise

We now consider the Lasso estimator based on n noisy observations. Let $X_i \in \mathcal{X}$ $(i = 1, \ldots, n)$ be the co-variables, and $Y_i \in \mathbb{R}$ $(i = 1, \ldots, n)$ be the response variables. The noisy Lasso is

$$\hat{\beta} := \arg\min_{\beta} \biggl\{ \frac{1}{n} \sum_{i=1}^n |Y_i - f_\beta(X_i)|^2 + \lambda \|\beta\|_1 \biggr\}.$$

The design matrix is

$$\mathbf{X} = \mathbf{X}_{n \times p} := (\psi_j(X_i)).$$

The empirical Gram matrix is

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X}/n = \int \psi^T \psi dQ_n = (\hat{\sigma}_{j,k}),$$

where Q_n is the empirical measure $Q_n := \sum_{i=1}^n \delta_{X_i}/n$. The $L_2(Q_n)$-norm is denoted by $\| \cdot \|_n$. We moreover let $(\cdot, \cdot)_n$ be the $L_2(Q_n)$-inner product.

As before, we write $f^0 = f_{\beta^0}$ and now, $\hat{f} = f_{\hat{\beta}}$. We consider

$$\epsilon_i := Y_i - f^0(X_i), \ i = 1, \ldots, n,$$

as the *noise*. Moreover, we write (with some abuse of notation)

$$(f, \epsilon)_n := \frac{1}{n} \sum_{i=1}^n f(X_i) \epsilon_i,$$

and we define

$$\lambda_0 := 2 \max_{1 \le j \le p} |(\psi_j, \epsilon)_n|.$$

Here is a simple example which shows how λ_0 behaves in the case of i.i.d. standard normal errors.

Lemma 11.1 *Suppose that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0,1)$-distributed, and that $\hat{\sigma}_{j,j} = 1$ for all j. Then we have for all $t > 0$, and for*

$$\lambda_0(t) := 2 \sqrt{\frac{2t + 2 \log p}{n}},$$

$$\mathbb{P}\biggl(2 \max_{1 \le j \le p} |(\psi_j, \epsilon)_n| \le \lambda_0(t) \biggr) \ge 1 - 2 \exp[-t].$$

Proof. As $\hat\sigma_{j,j} = 1$, we know that $V_j := \sqrt{n}(\psi_j, \epsilon)_n$ is $\mathcal{N}(0,1)$-distributed. So

$$\mathbb{P}\left(\max_{1\le j\le p} |V_j| > \sqrt{2t + 2\log p}\right) \le 2p \exp\left[-\frac{2t+2\log p}{2}\right] = 2\exp\left[-t\right].$$

$\square$

11.1 Prediction error in the noisy case

A noisy counterpart of Lemma 2.1 is:

Lemma 11.2 *Take $\lambda > \lambda_0$, and define $L := (\lambda + \lambda_0)/(\lambda - \lambda_0)$. Then*

$$\|\hat f - f^0\|_n^2 + \frac{2\lambda_0}{L-1}\|\hat\beta_{S^c}\|_1 \le \frac{4(L+1)^2\lambda_0^2 s}{(L-1)^2\phi^2_{\rm compatible}(\hat\Sigma, L, S)}.$$

Proof of Lemma 11.2. Because

$$2|(\epsilon, \hat f - f^0)| \le \left(2\max_{1\le j\le p}|(\psi_j,\epsilon)|\right)\|\hat\beta - \beta^0\|_1 \le \lambda_0\|\hat\beta - \beta^0\|_1,$$

we now have the Basic Inequality

$$\|\hat f - f^0\|_n^2 + \lambda\|\hat\beta_{S^c}\|_1 \le \lambda_0\|\hat\beta - \beta^0\|_1 + \lambda\|\beta^0\|_1.$$

Hence,

$$\|\hat f - f^0\|_n^2 + (\lambda - \lambda_0)\|\hat\beta_{S^c}\|_1 \le (\lambda + \lambda^0)\|\hat\beta_S - \beta_S^0\|_1.$$

Thus,

$$\|\hat\beta_{S^c}\|_1 \le L\|\hat\beta_S - \beta_S^0\|_1.$$

This implies

$$\|\hat\beta_S - \beta_S^0\|_1 \le \sqrt{s}\|\hat f - f^0\|_n/\phi_{\rm compatible}(\hat\Sigma, L, S).$$

So we arrive at

$$\|\hat f - f^0\|_n^2 + (\lambda - \lambda_0)\|\hat\beta_{S^c}\|_1 \le (\lambda + \lambda^0)\sqrt{s}\|\hat f - f^0\|_n/\phi_{\rm compatible}(\hat\Sigma, L, S).$$

Now, insert $\lambda = \lambda_0(L+1)/(L-1)$. $\square$

In a similar way, but using $(S, 2s)$-restricted eigenvalue conditions, one may prove ℓ_2-convergence in the noisy case.

Observe that the S-compatibility condition now involves the matrix $\hat\Sigma$, which is definitely singular when $p > n$. However, we have seen in the previous section that, also for such $\hat\Sigma$, compatibility conditions and restricted eigenvalue conditions hold in fairly general situations.

11.2 Noisy KKT

The KKT conditions in the noisy case become

$$2(\psi_j, \hat{f} - f^0)_n - 2(\psi_j, \epsilon)_n = -\lambda \hat{\tau}_j,\ j = 1, \ldots, p,$$

or in matrix notation,

$$2\hat{\Sigma}(\hat{\beta} - \beta^0) - \mathbf{X}^T \epsilon / n = -\lambda \hat{\tau},$$

where $\|\hat{\tau}\|_\infty \leq 1$, and $\hat{\tau}_j := \text{sign}(\hat{\beta}_j)$ whenever $\hat{\beta}_j \neq 0$.

To avoid too many repetitions, let us only formulate the noisy version of a part of Part 1 of Lemma 6.2.

Lemma 11.3 *Take $\lambda > \lambda_0$, and define $L := (\lambda + \lambda_0)/(\lambda - \lambda_0)$. Suppose the uniform $(\hat{\Sigma}, L, S, s)$-irrepresentable condition holds. Then $\hat{S} \subset S$.*

Proof of Lemma 11.3. This follows from a straightforward generalization of Lemma 6.1, where the equalities now become inequalities:

$$2\|(f_{\hat{\beta}_{S^c}})^{\hat{A}_S}\|_n^2 \leq \frac{2L}{L-1} \lambda_0 \hat{\Sigma}_{2,1}(S) \hat{\Sigma}_{1,1}^{-1}(S) \hat{\tau}_S - \frac{2}{L-1} \lambda_0 \|\hat{\beta}_{S^c}\|_1.$$

Here, $f^{\hat{A}_S}$ is the anti-projection of f, in $L_2(Q_n)$, on the space spanned by $\{\psi_j\}_{j \in S}$.

□

The noisy KKT conditions involve the matrix $\hat{\Sigma}$. Again, as discussed in Subsection 10.1, we may replace it by an approximation. As a consequence, if this approximation is good enough, we can replace $(\hat{\Sigma}, L, S, s)$-irrepresentable conditions by $(\Sigma, \tilde{L}, S, s)$-irrepresentable conditions, provided we take $\tilde{L} > L$ large enough.

Lemma 11.4 *Take $\lambda > \lambda_0$, and define $L := (\lambda + \lambda_0)/(\lambda - \lambda_0)$. Suppose that*

$$d_\infty(\hat{\Sigma}, \Sigma) \leq \tilde{\lambda},$$

and

$$\phi_{\text{compatible}}(\Sigma, L, S) > (L+1)\sqrt{\tilde{\lambda} s}$$

and in fact, that

$$\frac{(L+1)\sqrt{\tilde{\lambda} s}}{\phi_{\text{compatible}}(\Sigma, L, S) - (L+1)\sqrt{\tilde{\lambda} s}} < 1.$$

Then

$$\|(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0)\|_\infty < \frac{2\lambda_0}{L-1}.$$

Proof of Lemma 11.4. We have

$$\|(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0)\|_\infty \leq \tilde{\lambda} \|\hat{\beta} - \beta^0\|_1 \leq (L+1)\tilde{\lambda} \|\hat{\beta}_S - \beta_S^0\|_1$$

$$\leq (L+1)\tilde{\lambda}\sqrt{s} \|\hat{f} - f^0\|_n / \phi_{\text{compatible}}(\hat{\Sigma}, L, S)$$

$$\leq \frac{2\lambda_0 (L+1)^2 \tilde{\lambda} s}{(L-1)\phi^2_{\text{compatible}}(\hat{\Sigma}, L, S)}$$

$$\leq \frac{2\lambda_0 (L+1)^2 \tilde{\lambda} s \lambda_0}{(L-1)\left(\phi_{\text{compatible}}(\Sigma, L, S) - (L+1)\sqrt{\tilde{\lambda} s}\right)^2}.$$

□

We conclude that the KKT conditions in the noisy case can be exploited in the same way as in the case without noise, albeit that one needs to adjust the constants (making the conditions more restrictive).

12 Discussion

We show how various conditions for Lasso oracle results relate to each other, as illustrated in Figure 1. Thereby, we also introduce the restricted regression condition.

For deriving oracle results for prediction and estimation, the compatibility condition is the weakest. Looking at the derivation of the oracle result in Lemma 2.1, no substantial room seems to be left to improve the condition. The restricted eigenvalue condition is slightly stronger but in some cases, as demonstrated in Example 10.5, the compatibility condition is a real improvement.

For variable selection with the Lasso, the irrepresentable condition is sufficient (assuming sufficiently large non-zero regression coefficients) and essentially necessary. We present the, perhaps not unexpected, but as yet not formally shown, result that the irrepresentable condition is always stronger than the compatibility condition.

We illustrate in Section 10 how - in theory - one can verify the compatibility condition. If the sparsity is of small order $o(\sqrt{n/\log p})$, we can approximate the empirical Gram matrix by the population analogue. It is then much more easy and realistic that the population Gram matrix has sufficiently regular behavior, as illustrated with our examples in Section 10.2. We believe moreover that a sparsity bound of small order $o(\sqrt{n/\log p})$ covers a large area of interesting statistical problems. With larger s, the statistical situation is comparable to one of a nonparametric model with "(effective) smoothness less than 1/2", leading to very slow convergence rates. In contrast, for example in decoding problems, sparseness up to the linear-in-n regime can be very important. Moreover, in the case of robust convex loss, one may apply the compatibility condition directly to the population matrix, i.e., the sparsity regime $s = o(\sqrt{n/\log p})$ can be relaxed for such loss functions (see van de Geer (2008)). We therefore conclude that oracle results for the Lasso hold under quite general design conditions.

A final remark is that in our formulation, the compatibility condition and restricted eigenvalue condition depend on the sparsity s as well as on the active set S. As S is unknown, this means that for a practical guarantee, the conditions should hold for *all* S. Moreover, one then needs to assume the sparsity s to be known, or at least a good upper bound needs to be given. Such strong requirements are the price for practical verifiability. We

however believe that in statistical modeling, non-verifiable conditions are allowed and in fact common practice. Moreover, our model assumes a sparse linear "truth" with "true" active set S, only for simplicity. Without such assumptions, there is no "true" S, and the oracle inequality concerns a trade-off between sparse approximation and estimation error, see for example van de Geer (2008).

References

D. Bertsimas and J. Tsitsiklis. *Introduction to linear optimization.* Athena Scientific Belmont, MA, 1997.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674, 2007a.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparse Density Estimation with ℓ_1 Penalties. In *Learning Theory 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA, June 13-15, 2007: Proceedings*, page 530. Springer, 2007b.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007c.

T. Cai, G. Xu, and J. Zhang. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55:3388–3397, 2009.

T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *Preprint*, 2009a.

T. Cai, L. Wang, and G. Xu. Stable recovery of sparse signals and an oracle inequality. *Preprint*, 2009b.

E. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37:2145–2177, 2009.

E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45:7–57, 2009a.

V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15: 799–828, 2009b.

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270, 2009.

S. Parter. Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations. *Transactions of the American Mathematical Society*, 99:153–192, 1961.

S. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645, 2008.

S. van de Geer. The deterministic Lasso. In *JSM proceedings,* (see also http://stat.ethz.ch/research/research_reports/2007/140). American Statistical Association, 2007.

M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

T. Zhang. Some sharp performance bounds for least squares regression with L1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.